

PATENT COOPERATION TREATY

PCT

From the INTERNATIONAL BUREAU

To:

Assistant Commissioner for Patents
 United States Patent and Trademark
 Office
 Box PCT
 Washington, D.C.20231
 ETATS-UNIS D'AMERIQUE

in its capacity as elected Office

Date of mailing (day/month/year) 12 May 2000 (12.05.00)	Applicant's or agent's file reference GR 98P2663P
International application No. PCT/DE99/02846	Priority date (day/month/year) 23 September 1998 (23.09.98)
International filing date (day/month/year) 08 September 1999 (08.09.99)	

Applicant NEUNEIER, Ralf et al

1. The designated Office is hereby notified of its election made:

in the demand filed with the International Preliminary Examining Authority on:

11 April 2000 (11.04.00)

in a notice effecting later election filed with the International Bureau on:

2. The election was was not

made before the expiration of 19 months from the priority date or, where Rule 32 applies, within the time limit under Rule 32.2(b).

The International Bureau of WIPO 34, chemin des Colombettes 1211 Geneva 20, Switzerland	Authorized officer Kiwa Mpay
Facsimile No.: (41-22) 740.14.35	Telephone No.: (41-22) 338.83.38

6-T

**VERTRAG ÜBER DIE INTERNATIONALE ZUSAMMENARBEIT AUF DEM
GEBIET DES PATENTWESENS**

PCT

REC'D 22 DEC 2000

PCT

INTERNATIONALER VORLÄUFIGER PRÜFUNGSBERICHT

(Artikel 36 und Regel 70 PCT)

Aktenzeichen des Anmelders oder Anwalts GR 98P2663P	WEITERES VORGEHEN	siehe Mitteilung über die Übersendung des internationalen vorläufigen Prüfungsberichts (Formblatt PCT/IPEA/416)
Internationales Aktenzeichen PCT/DE99/02846	Internationales Anmeldedatum (Tag/Monat/Jahr) 08/09/1999	Prioritätsdatum (Tag/Monat/Tag) 23/09/1998
Internationale Patentklassifikation (IPK) oder nationale Klassifikation und IPK G06F17/60		
Anmelder SIEMENS AKTIENGESELLSCHAFT et al.		
<p>1. Dieser internationale vorläufige Prüfungsbericht wurde von der mit der internationalen vorläufigen Prüfung beauftragten Behörde erstellt und wird dem Anmelder gemäß Artikel 36 übermittelt.</p> <p>2. Dieser BERICHT umfaßt insgesamt 4 Blätter einschließlich dieses Deckblatts.</p> <p><input type="checkbox"/> Außerdem liegen dem Bericht ANLAGEN bei; dabei handelt es sich um Blätter mit Beschreibungen, Ansprüchen und/oder Zeichnungen, die geändert wurden und diesem Bericht zugrunde liegen, und/oder Blätter mit vor dieser Behörde vorgenommenen Berichtigungen (siehe Regel 70.16 und Abschnitt 607 der Verwaltungsrichtlinien zum PCT).</p> <p>Diese Anlagen umfassen insgesamt Blätter.</p>		
<p>3. Dieser Bericht enthält Angaben zu folgenden Punkten:</p> <ul style="list-style-type: none"> I <input checked="" type="checkbox"/> Grundlage des Berichts II <input type="checkbox"/> Priorität III <input type="checkbox"/> Keine Erstellung eines Gutachtens über Neuheit, erforderliche Tätigkeit und gewerbliche Anwendbarkeit IV <input type="checkbox"/> Mangelnde Einheitlichkeit der Erfindung V <input checked="" type="checkbox"/> Begründete Feststellung nach Artikel 35(2) hinsichtlich der Neuheit, der erforderlichen Tätigkeit und der gewerblichen Anwendbarkeit; Unterlagen und Erklärungen zur Stützung dieser Feststellung VI <input type="checkbox"/> Bestimmte angeführte Unterlagen VII <input checked="" type="checkbox"/> Bestimmte Mängel der internationalen Anmeldung VIII <input type="checkbox"/> Bestimmte Bemerkungen zur internationalen Anmeldung 		

Datum der Einreichung des Antrags 11/04/2000	Datum der Fertigstellung dieses Berichts 21.12.2000
Name und Postanschrift der mit der internationalen vorläufigen Prüfung beauftragten Behörde:  Europäisches Patentamt - P.B. 5818 Patentlaan 2 NL-2280 HV Rijswijk - Pays Bas Tel. +31 70 340 - 2040 Tx: 31 651 epo nl Fax: +31 70 340 - 3016	Bevollmächtigter Bediensteter Skulikaris, I Tel. Nr. +31 70 340 3185



**INTERNATIONALER VORLÄUFIGER
PRÜFUNGSBERICHT**

Internationales Aktenzeichen PCT/DE99/02846

I. Grundlage des Berichts

1. Dieser Bericht wurde erstellt auf der Grundlage (*Ersatzblätter, die dem Anmeldeamt auf eine Aufforderung nach Artikel 14 hin vorgelegt wurden, gelten im Rahmen dieses Berichts als "ursprünglich eingereicht" und sind ihm nicht beigefügt, weil sie keine Änderungen enthalten.:*)

Beschreibung, Seiten:

1-28 ursprüngliche Fassung

Patentansprüche, Nr.:

1-21 ursprüngliche Fassung

Zeichnungen, Blätter:

1-4 ursprüngliche Fassung

2. Hinsichtlich der **Sprache**: Alle vorstehend genannten Bestandteile standen der Behörde in der Sprache, in der die internationale Anmeldung eingereicht worden ist, zur Verfügung oder wurden in dieser eingereicht, sofern unter diesem Punkt nichts anderes angegeben ist.

Die Bestandteile standen der Behörde in der Sprache: zur Verfügung bzw. wurden in dieser Sprache eingereicht; dabei handelt es sich um

- die Sprache der Übersetzung, die für die Zwecke der internationalen Recherche eingereicht worden ist (nach Regel 23.1(b)).
- die Veröffentlichungssprache der internationalen Anmeldung (nach Regel 48.3(b)).
- die Sprache der Übersetzung, die für die Zwecke der internationalen vorläufigen Prüfung eingereicht worden ist (nach Regel 55.2 und/oder 55.3).

3. Hinsichtlich der in der internationalen Anmeldung offenbarten **Nucleotid- und/oder Aminosäuresequenz** ist die internationale vorläufige Prüfung auf der Grundlage des Sequenzprotokolls durchgeführt worden, das:

- in der internationalen Anmeldung in schriftlicher Form enthalten ist.
- zusammen mit der internationalen Anmeldung in computerlesbarer Form eingereicht worden ist.
- bei der Behörde nachträglich in schriftlicher Form eingereicht worden ist.
- bei der Behörde nachträglich in computerlesbarer Form eingereicht worden ist.
- Die Erklärung, daß das nachträglich eingereichte schriftliche Sequenzprotokoll nicht über den Offenbarungsgehalt der internationalen Anmeldung im Anmeldezeitpunkt hinausgeht, wurde vorgelegt.
- Die Erklärung, daß die in computerlesbarer Form erfassten Informationen dem schriftlichen Sequenzprotokoll entsprechen, wurde vorgelegt.

4. Aufgrund der Änderungen sind folgende Unterlagen fortgefallen:

INTERNATIONALER VORLÄUFIGER PRÜFUNGSBERICHT

Internationales Aktenzeichen PCT/DE99/02846

Beschreibung, Seiten:
 Ansprüche, Nr.:
 Zeichnungen, Blatt:

5. Dieser Bericht ist ohne Berücksichtigung (von einigen) der Änderungen erstellt worden, da diese aus den angegebenen Gründen nach Auffassung der Behörde über den Offenbarungsgehalt in der ursprünglich eingereichten Fassung hinausgehen (Regel 70.2(c)).

(Auf Ersatzblätter, die solche Änderungen enthalten, ist unter Punkt 1 hinzuweisen; sie sind diesem Bericht beizufügen).

6. Etwaige zusätzliche Bemerkungen:

V. Begründete Feststellung nach Artikel 35(2) hinsichtlich der Neuheit, der erfinderischen Tätigkeit und der gewerblichen Anwendbarkeit; Unterlagen und Erklärungen zur Stützung dieser Feststellung

1. Feststellung

Neuheit (N)	Ja: Ansprüche	1-21
	Nein: Ansprüche	
Erfinderische Tätigkeit (ET)	Ja: Ansprüche	1-21
	Nein: Ansprüche	
Gewerbliche Anwendbarkeit (GA)	Ja: Ansprüche	1-21
	Nein: Ansprüche	

2. Unterlagen und Erklärungen siehe Beiblatt

VII. Bestimmte Mängel der internationalen Anmeldung

Es wurde festgestellt, daß die internationale Anmeldung nach Form oder Inhalt folgende Mängel aufweist:
siehe Beiblatt

Zu Punkt V:

- 1 Die in dem Internationalen Recherchenbericht zitierten Dokumente in der darin enthaltenen Reihenfolge werden als D1, D2 und D3 bezeichnet.
- 2 D1 offenbart ein Verfahren zur Ermittlung einer Folge von Aktionen für ein System, welches Zustände aufweist, wobei ein Zustandsübergang zwischen zwei Zuständen aufgrund einer Aktion erfolgt.
- 3 D2 offenbart ein Verfahren im Kontext von Softwarebenutzung, siehe Abstract, ebenfalls Teil 1, Introduction. D2 offenbart insbesondere die Berücksichtigung eines Risikos (failure risk, Seite 185, 2. Absatz), sowie die Berücksichtigung eines einstellbaren Parameters (usage distribution, i.e. the frequency of invocation of critical operations) zum Zweck der Optimierung. Die Zielsetzung bei D2 ist jedoch, die Übergangswahrscheinlichkeiten für Systemübergänge zu ändern, so daß kritische Systemoperationen oft genug getestet werden, und die Risikoschätzung mit höchstmöglicher Präzision stattfindet. D2 betrifft nicht die Ermittlung einer Folge von Aktionen. Auch aus D1 erhält der Fachmann keinen Hinweis, ein Dokument mit dem Inhalt von D2 zu konsultieren. Demgemäß wäre es wegen der unterschiedlichen Zielsetzung fraglich, ob der Fachmann das Dokument D2 hinzuziehen würde.
- 4 Aus diesen Gründen erfüllen die Ansprüche 1 und 15 die Erfordernisse des Artikels 33 PCT. Diese Schlußfolgerung gilt ebenfalls für die abhängigen Ansprüche.

Zu Punkt VII:

- 1 Die gegenwärtige Formulierung der Ansprüche 1-6 läßt erkennen, daß der Gegenstand dieser Ansprüche u.U. unter die Bestimmungen des Artikels 34(4)(a)(i) und der Regel 67.1(i) PCT (mathematische Verfahren) fällt.

091787698

Translation

PATENT COOPERATION TREATY

PCT

3

INTERNATIONAL PRELIMINARY EXAMINATION REPORT

(PCT Article 36 and Rule 70)

Applicant's or agent's file reference GR 98P2663P	FOR FURTHER ACTION	SeeNotificationofTransmittalofInternational Preliminary Examination Report (Form PCT/IPEA/416)
International application No. PCT/DE99/02846	International filing date (day/month/year) 08 September 1999 (08.09.99)	Priority date (day/month/year) 23 September 1998 (23.09.98)
International Patent Classification (IPC) or national classification and IPC G06F 17/60		
Applicant SIEMENS AKTIENGESELLSCHAFT		

1. This international preliminary examination report has been prepared by this International Preliminary Examining Authority and is transmitted to the applicant according to Article 36.
2. This REPORT consists of a total of 4 sheets, including this cover sheet.

This report is also accompanied by ANNEXES, i.e., sheets of the description, claims and/or drawings which have been amended and are the basis for this report and/or sheets containing rectifications made before this Authority (see Rule 70.16 and Section 607 of the Administrative Instructions under the PCT).

These annexes consist of a total of _____ sheets.

3. This report contains indications relating to the following items:

- I Basis of the report
- II Priority
- III Non-establishment of opinion with regard to novelty, inventive step and industrial applicability
- IV Lack of unity of invention
- V Reasoned statement under Article 35(2) with regard to novelty, inventive step or industrial applicability; citations and explanations supporting such statement
- VI Certain documents cited
- VII Certain defects in the international application
- VIII Certain observations on the international application

RECEIVED

JULY 02 2001

Technology Center 2100

Date of submission of the demand 11 April 2000 (11.04.00)	Date of completion of this report 21 December 2000 (21.12.2000)
Name and mailing address of the IPEA/EP	Authorized officer
Facsimile No.	Telephone No.

INTERNATIONAL PRELIMINARY EXAMINATION REPORT

International application No.

PCT/DE99/02846

I. Basis of the report

1. With regard to the elements of the international application:*

 the international application as originally filed the description:

pages 1-28, as originally filed

pages _____, filed with the demand

pages _____, filed with the letter of _____

 the claims:

pages 1-21, as originally filed

pages _____, as amended (together with any statement under Article 19)

pages _____, filed with the demand

pages _____, filed with the letter of _____

 the drawings:

pages 1-4, as originally filed

pages _____, filed with the demand

pages _____, filed with the letter of _____

 the sequence listing part of the description:

pages _____, as originally filed

pages _____, filed with the demand

pages _____, filed with the letter of _____

2. With regard to the language, all the elements marked above were available or furnished to this Authority in the language in which the international application was filed, unless otherwise indicated under this item.

These elements were available or furnished to this Authority in the following language _____ which is:

 the language of a translation furnished for the purposes of international search (under Rule 23.1(b)). the language of publication of the international application (under Rule 48.3(b)). the language of the translation furnished for the purposes of international preliminary examination (under Rule 55.2 and/or 55.3).

3. With regard to any nucleotide and/or amino acid sequence disclosed in the international application, the international preliminary examination was carried out on the basis of the sequence listing:

 contained in the international application in written form. filed together with the international application in computer readable form. furnished subsequently to this Authority in written form. furnished subsequently to this Authority in computer readable form. The statement that the subsequently furnished written sequence listing does not go beyond the disclosure in the international application as filed has been furnished. The statement that the information recorded in computer readable form is identical to the written sequence listing has been furnished.4. The amendments have resulted in the cancellation of: the description, pages _____ the claims, Nos. _____ the drawings, sheets/fig. _____5. This report has been established as if (some of) the amendments had not been made, since they have been considered to go beyond the disclosure as filed, as indicated in the Supplemental Box (Rule 70.2(c)).**

* Replacement sheets which have been furnished to the receiving Office in response to an invitation under Article 14 are referred to in this report as "originally filed" and are not annexed to this report since they do not contain amendments (Rule 70.16 and 70.17).

** Any replacement sheet containing such amendments must be referred to under item 1 and annexed to this report.

INTERNATIONAL PRELIMINARY EXAMINATION REPORTInternational application No.
PCT/DE 99/02846**V. Reasoned statement under Article 35(2) with regard to novelty, inventive step or industrial applicability: citations and explanations supporting such statement****1. Statement**

Novelty (N)	Claims	1-21	YES
	Claims		NO
Inventive step (IS)	Claims	1-21	YES
	Claims		NO
Industrial applicability (IA)	Claims	1-21	YES
	Claims		NO

2. Citations and explanations

- 1 The documents cited in the international search report are referred to as D1, D2 and D3, in the order of their listing therein.
- 2 D1 discloses a method for the determination of a sequence of actions for a system which comprises statuses, with a status transition between two statuses taking place because of an action.
- 3 D2 discloses a method in the context of software use - see abstract and also part 1, Introduction. D2 discloses in particular the consideration of a risk (failure risk, page 185, second paragraph) and the consideration of an adjustable parameter (usage distribution, i.e. the frequency of invocation of critical operations) for optimisation. The objective in D2, however, is to change the transition probabilities of system transitions, and hence critical system operations are tested often enough and that a risk estimation of the highest possible precision takes place. D2 does not apply to the determination of a sequence of actions. Neither does D1 suggest to a person skilled in the art to consult

INTERNATIONAL PRELIMINARY EXAMINATION REPORT

International application No.

PCT/DE 99/02846

different objectives it would therefore be doubtful whether the person skilled in the art would refer to D2.

4 Because of these reasons Claims 1 and 15 meet the requirements of PCT Article 33. This conclusion also applies to the dependent claims.

RECEIVED
JUL 02 2001
Technology Center 2100

VII. Certain defects in the international application

The following defects in the form or contents of the international application have been noted:

- 1 By the present wording of Claims 1-6 the subject matter of these claims may possibly come under the terms of PCT Article 33(4)(a)(i) and PCT Rule 67.1 (mathematical theories).

Beschreibung

5 **Verfahren und Anordnung zur Ermittlung einer Folge von Aktionen für ein System, welches Zustände aufweist, wobei ein Zustandsübergang zwischen zwei Zuständen aufgrund einer Aktion erfolgt**

10 Die Erfindung betrifft ein Verfahren sowie eine Anordnung zur Ermittlung einer Folge von Aktionen für ein System, welches Zustände aufweist, wobei ein Zustandsübergang zwischen zwei Zuständen aufgrund einer Aktion erfolgt.

15 Ein solches Verfahren und eine solche Anordnung sind aus [1] bekannt.

20 In [1] ist als Beispiel für ein solches System, welches Zustände aufweist, ein Finanzmarkt beschrieben.

25 Das System wird als ein Markov-Entscheidungsproblem beschrieben (Markov-Decision-Problem, MDP). Ein System, welches als Markov-Entscheidungsproblem beschrieben werden kann, ist in seiner Struktur in Fig.2 dargestellt.

30 Zu einem Zeitpunkt t befindet sich das System 201 in einem Zustand x_t . Der Zustand x_t ist für einen Beobachter des Systems beobachtbar. Aufgrund einer Aktion a_t aus einer Menge in dem Zustand x_t möglicher Aktionen, $a_t \in A(x_t)$ geht das System mit einer gewissen Wahrscheinlichkeit in einen Folgezustand x_{t+1} zu einem Folgezeitpunkt $t+1$ über.

35 Dies ist durch eine Schleife in Fig.2 symbolisch dargestellt. Ein Beobachter 200 nimmt beobachtbare Größen über den Zustand x_t wahr 202 und trifft eine Entscheidung über eine Aktion 203, mit der er auf das System 201 einwirkt. Das System 201 unterliegt üblicherweise einer Störung 205.

Ferner erhält der Beobachter 200 einen Gewinn r_t 204

$$r_t = r(x_t, a_t, x_{t+1}) \in \mathbb{R}, \quad (1)$$

der von der Aktion a_t 203 und dem ursprünglichen Zustand x_t 5 zu dem Zeitpunkt t sowie dem Folgezustand x_{t+1} des Systems zu dem Folgezeitpunkt $t+1$ abhängt.

Der Gewinn r_t kann einen positiven oder negativen skalaren Wert annehmen, je nachdem, ob die Entscheidung zu einer 10 hinsichtlich eines vorgebbaren Kriteriums positiven oder negativen Systementwicklung führt, in [1] zu einer Kapitalvermehrung oder zu einem Verlust.

In einem weiteren Zeitschritt entscheidet sich der Beobachter 15 200 des Systems 201 aufgrund der beobachtbaren Größen 202, 204 des Folgezustandes x_{t+1} für eine neue Aktion a_{t+1} usw.

Eine Folge von

$$\begin{array}{llll} \text{Zustand:} & x_t & \in & X \\ \text{Aktion:} & a_t & \in & A(x_t) \\ \text{Folgezustand:} & x_{t+1} & \in & X \\ \text{Gewinn} & r_t = r(x_t, a_t, x_{t+1}) & \in & \mathbb{R} \end{array}$$

20 usw. beschreibt eine Trajektorie des Systems, die durch ein Performanzkriterium, das die einzelnen Gewinne r_t über die Zeitpunkte t akkumuliert, bewertet wird. Bei einem Markov-Entscheidungsproblem wird vereinfachend angenommen, daß der 25 Zustand x_t und die Aktion a_t alle Informationen enthalten, um eine Übergangswahrscheinlichkeit $p(x_{t+1}| \cdot)$ des Systems von dem Zustand x_t zu dem Folgezustand x_{t+1} zu beschreiben.

Formal bedeutet dies:

$$p(x_{t+1}|x_t, K, x_0, a_t, K, a_0) = p(x_{t+1}|x_t, a_t). \quad (2)$$

Mit $p(x_{t+1}|x_t, a_t)$ wird eine Übergangswahrscheinlichkeit für den Folgezustand x_{t+1} bei gegebenem Zustand x_t und gegebener Aktion a_t bezeichnet.

5 Bei einem Markov-Entscheidungsproblem hängen also zukünftige Zustände des Systems 201 nicht von Zuständen und Aktionen ab, die weiter als einen Zeitschritt in der Vergangenheit liegen.

10 Zusammenfassend sind im weiteren die Charakteristika eines Markov-Entscheidungsproblems dargestellt:

X	Menge der möglichen Zustände des Systems, z.B. $X = \mathbb{R}^m$,
$A(x_t)$	Menge der möglichen Aktionen in dem Zustand
$p(x_{t+1} x_t, a_t)$	x_t
$r(x_t, a_t, x_{t+1})$	Gewinn mit Erwartungswert $R(x_t, a_t)$.

Das Ziel ist es, ausgehend von beobachtbaren Größen, den im weiteren als Trainingsdaten bezeichneten Größen, eine Strategie zu ermitteln, d.h. eine Folge von Funktionen

$$\pi = \{\mu_0, \mu_1, \dots, \mu_T\}, \quad (3)$$

20 welche zu jedem Zeitpunkt t jeden Zustand in eine Handlungsvorschrift, d.h. Aktion

$$\mu_t(x_t) = a_t \quad (4)$$

abbilden.

25 Eine solche Strategie wird durch eine Optimierungsfunktion bewertet.

Die Optimierungsfunktion gibt den Erwartungswert, der über die Zeit akkumulierten Gewinne bei einer gegebenen Strategie π und einem Startzustand x_0 an.

5 Als ein Beispiel eines Verfahrens des approximativen dynamischen Programmierens ist in [1] das sogenannte Q-Lernverfahren beschrieben.

Eine optimale Bewertungsfunktion $V^*(x)$ ist definiert durch
10

$$V^*(x) = \max_{\pi} V^{\pi}(x) \quad \forall x \in X \quad (5)$$

mit

15 $V^{\pi}(x) = E \left[\sum_{t=0}^{\infty} \gamma^t r(x_t, \mu_t, x_{t+1}) | x_0 = x \right], \quad (6)$

wobei mit γ ein vorgebbarer Verringerungsfaktor bezeichnet wird, der gemäß folgender Vorschrift gebildet wird:

20 $\gamma = \frac{1}{1+z}, \quad (7)$

$z \in \mathbb{R}^+.$ (8)

Im Rahmen des Q-Lernverfahrens wird für jedes Paar
25 (Zustand x_t , Aktion a_t) eine Q-Bewertungsfunktion $Q^*(x_t, a_t)$ gemäß folgender Vorschrift gebildet:

$$Q^*(x_t, a_t) := \sum_{x \in X} p(x_{t+1} | x_t, a_t) \cdot r_t + \\ + \gamma \cdot \sum_{x \in X} p(x | x_t, a_t) \cdot \max_{a \in A} (Q^*(x, a))$$

(9)

Aufgrund jeweils des Tupels (x_t, x_{t+1}, a_t, r_t) werden die Q-Werte $Q^*(x, a)$ in der $k+1$ ten Iteration gemäß folgender Lernregel mit einer vorgegebenen Lernrate η_k gemäß folgender Vorschrift adaptiert:

5

$$Q_{k+1}(x_t, a_t) = (1 - \eta_k)Q_k(x_t, a_t) + \eta_k \left(r_t + \gamma \max_{a \in A} (Q_k(x_{t+1}, a)) \right). \quad (10)$$

Üblicherweise werden die sogenannten Q-Werte $Q^*(x, a)$ durch jeweils einen Funktionsapproximator, beispielsweise ein neuronales Netz oder auch einen Polynomklassifikator, mit einem Gewichtsvektor w^a , der Gewichte des Funktionsapproximators enthält, für verschiedene Aktionen a approximiert.

Unter einem Funktionsapproximator ist beispielsweise ein neuronales Netz, ein Polynomklassifikator oder auch eine Kombination eines neuronalen Netzes mit einem Polynomklassifikator zu verstehen.

Es gilt also:

$$Q^*(x, a) \approx Q(x; w^a). \quad (11)$$

Änderungen der Gewichte in dem Gewichtsvektor w^a basieren auf einer temporären Differenz d_t , die gemäß folgender Vorschrift gebildet wird:

$$d_t := r(x_t, a_t, x_{t+1}) + \gamma \max_{a \in A} Q(x_{t+1}; w_k^a) - Q(x_t; w_k^{a_t}) \quad (12)$$

Es ergibt sich für das Q-Lernverfahren unter Verwendung eines neuronalen Netzes folgende Adoptionsvorschrift für die Gewichte des neuronalen Netzes, welche Gewichte in dem Gewichtsvektor w^a enthalten sind:

$$w_{k+1}^{at} = w_k^{at} + \eta_k \cdot d_t \cdot \nabla Q(x_t; w_k^{at}). \quad (13)$$

Unter Verwendung der Trainingsdaten, die als Zeitreihenwerte 5 Informationen über vorangegangene Kursverläufe eines Finanzmarktes beschreiben, wird das neuronale Netz, welches das System Finanzmarkt, wie in [1] beschrieben, darstellt, trainiert.

Ein weiteres Verfahren der approximativen dynamischen Programmierung, das sogenannte $TD(\lambda)$ -Lernverfahren, ist aus [2] 10 bekannt und wird im Zusammenhang mit einem Ausführungsbeispiel näher erläutert.

Ferner ist aus [3] bekannt, welches Risiko mit einer Strategie π und einem Ausgangszustand x_t verbunden ist. Ein Verfahren zur Risikovermeidung ist ebenfalls aus [3] bekannt. 15

Bei dem aus [3] bekannten Verfahren wird folgende Optimierungsfunktion, welche auch als erweiterte Q-Funktion 20 bezeichnet wird, verwendet:

maximiere

$$\left\{ \underline{Q}^\pi(x_t, a_t) = r(x_t, a_t, x_{t+1}) + \inf_{\substack{x_0, x_1, K \\ p(x_0, x_1, K) > 0}} \left\{ \sum_{k=1}^{\infty} \gamma^k r(x_k, \pi(x_k), x_{k+1}) \right\} \right\} \quad (14)$$

25

Die erweiterte Q-Funktion $\underline{Q}^\pi(x_t, a_t)$ beschreibt den schlechtesten Fall, falls in dem Zustand x_t die Aktion a_t ausgeführt wird und die Strategie π daraufhin verfolgt wird.

30 Die Optimierungsfunktion $\underline{Q}^\pi(x_t, a_t)$ für

$$\underline{Q}^*(x_t, a_t) := \max_{\pi \in \Pi} \underline{Q}^\pi(x_t, a_t)$$

(15)

ist gegeben, durch folgende Vorschrift:

5

$$\underline{Q}^*(x_t, a_t) = \min_{\substack{x \in X \\ P(x_{t+1}|x_t, a_t) > 0}} \left(r(x_t, a_t, x) + \gamma \cdot \max_{a \in A} \underline{Q}^*(x, a) \right). \quad (16)$$

Ein erheblicher Nachteil dieser Vorgehensweise ist darin zu sehen, daß nur der schlechteste Fall im Rahmen der Strategiefindung berücksichtigt wird. Dies spiegelt jedoch die Anforderungen verschiedenster technischer Systeme nur in unzureichendem Ausmaß wieder.

Aus [4] ist es ferner bekannt, eine Zugangskontrolle für ein Kommunikationsnetz sowie das Routing innerhalb des Kommunikationsnetzes als ein Problem der dynamischen Programmierung zu formulieren.

Somit liegt der Erfindung das Problem zugrunde, ein Verfahren sowie eine Anordnung zur Ermittlung einer Folge von Aktionen für ein System anzugeben, bei dem bzw. bei der eine erhöhte Flexibilität bei der Ermittlung der Strategie erreicht wird.

Das Problem wird durch das Verfahren sowie durch die Anordnung gemäß den Merkmalen der unabhängigen Patentansprüche gelöst.

Bei einem Verfahren zur rechnergestützten Ermittlung einer Folge von Aktionen für ein System, welches Zustände aufweist, wobei ein Zustandsübergang zwischen zwei Zuständen aufgrund einer Aktion erfolgt, erfolgt die Ermittlung der Folge von Aktionen derart, daß eine aus der Folge von Aktionen resultierende Folge von Zuständen hinsichtlich einer vorgegebenen Optimierungsfunktion optimiert ist, wobei die Optimierungs-

funktion einen variablen Parameter enthält, mit dem ein Risiko, welches die resultierende Folge von Zuständen hinsichtlich eines vorgegebenen Zustandes des Systems aufweist, einstellbar ist.

5

Eine Anordnung zur Ermittlung einer Folge von Aktionen für ein System, welches Zustände aufweist, wobei ein Zustandsübergang zwischen zwei Zuständen aufgrund einer Aktion erfolgt, weist einen Prozessor auf, der derart eingerichtet ist, daß die Ermittlung der Folge von Aktionen derart erfolgen kann, daß eine aus der Folge von Aktionen resultierende Folge von Zuständen hinsichtlich einer vorgegebenen Optimierungsfunktion optimiert ist, wobei die Optimierungsfunktion einen variablen Parameter enthält, mit dem ein Risiko, welches die resultierende Folge von Zuständen hinsichtlich eines vorgegebenen Zustandes des Systems aufweist, einstellbar ist.

Durch die Erfindung wird es erstmals möglich, in frei vorgebarer Genauigkeit im Rahmen einer Strategiefindung für eine möglichen Regelung oder Steuerung, allgemein einer Beeinflussung des Systems, ein Verfahren zur Ermittlung einer Folge von Aktionen anzugeben.

Bevorzugte Weiterbildungen der Erfindung ergeben sich aus den abhängigen Ansprüchen.

Die im weiteren beschriebenen Weiterbildungen gelten sowohl für das Verfahren als auch die Anordnung, wobei bei der Weiterbildung der Anordnung jeweils der Prozessor derart eingerichtet ist, daß die Weiterbildung realisierbar ist.

In einer bevorzugten Ausgestaltung wird zur Ermittlung ein Verfahren des approximativen dynamischen Programmierens eingesetzt, beispielsweise ein auf dem Q-Lernen basierendes Verfahren oder auch ein auf dem TD(λ)-Lernen basierendes Verfahren.

Im Rahmen des Q-Lernens wird bevorzugt die Optimierungsfunktion OFQ gemäß folgender Vorschrift gebildet:

$$OFQ = Q(x; w^a),$$

5

wobei mit

- x ein Zustand in einem Zustandsraum X ,
- a eine Aktion aus einem Aktionsraum A ,
- w^a die zur Aktion a gehörigen Gewichte eines Funktionsapproximators

bezeichnet wird/werden.

15 Im Rahmen des Q-Lernens wird zur Ermittlung der optimalen Gewichte w^a des Funktionsapproximators folgender Adoptions-
schritt ausgeführt:

$$w_{t+1}^{at} = w_t^{at} + \eta_t \cdot \mathbf{x}^k(d_t) \cdot \nabla Q(x_t; w_t^{at})$$

20

mit der Abkürzung

$$d_t = r(x_t, a_t, x_{t+1}) + \gamma \max_{a \in A} Q(x_{t+1}, w_t^a) - Q(x_t, w_t^{at})$$

25 wobei mit

- x_t, x_{t+1} jeweils ein Zustand in dem Zustandsraum X ,
- a_t eine Aktion aus einem Aktionsraum A ,
- γ ein vorgebbarer Verringerungsfaktor,
- 30 • w_t^{at} der zur Aktion a_t gehörige Gewichtsvektor vor dem Adaptionsschritt,
- w_{t+1}^{at} der zur Aktion a_t gehörige Gewichtsvektor nach dem Adaptionsschritt,
- η_t ($t = 1, \dots$) eine vorgebbare Schrittweitenfolge,

- $\kappa \in [-1; 1]$ ein Risikokontrollparameter,
- κ^k eine Risikokontrollfunktion $\kappa^k(\xi) = (1 - \kappa \text{sign}(\xi))\xi$,
- $\nabla Q(\cdot; \cdot)$ die Ableitung des Funktionsapproximators nach seinen Gewichten,
- 5 • $r(x_t, a_t, x_{t+1})$ ein Gewinn beim Zustandsübergang von dem Zustand x_t nach dem Folgezustand x_{t+1} ,

bezeichnet wird/werden.

10 Im Rahmen des $TD(\lambda)$ -Lernverfahrens wird die Optimierungsfunktion bevorzugt gemäß folgender Vorschrift gebildet:

$$OFTD = J(x; w)$$

15 wobei mit

- x ein Zustand in einem Zustandsraum X ,
- a eine Aktion aus einem Aktionsraum A ,
- w die Gewichte eines Funktionsapproximators

20 bezeichnet wird/werden.

Im Rahmen des $TD(\lambda)$ -Lernens wird zur Ermittlung der optimalen Gewichte w des Funktionsapproximators folgender Adoptions-
25 schritt ausgeführt:

$$w_{t+1} = w_t + \eta_t \cdot \kappa^k(d_t) \cdot z_t$$

mit den Abkürzungen

$$d_t = r(w_t, a_t, x_{t+1}) + \gamma J(x_{t+1}; w_t) - J(x_t; w_t),$$

$$z_t = \lambda \cdot \gamma \cdot z_{t-1} + \nabla J(x_t; w_t),$$

$$35 z_{-1} = 0,$$

wobei mit

- x_t, x_{t+1} jeweils ein Zustand in dem Zustandsraum X ,
- a_t eine Aktion aus einem Aktionsraum A ,
- 5 • γ ein vorgebbarer Verringerungsfaktor,
- w_t der Gewichtsvektor vor dem Adaptionsschritt,
- w_{t+1} der Gewichtsvektor nach dem Adaptionsschritt,
- η_t ($t = 1, \dots$) eine vorgebbare Schrittweitenfolge,
- $\kappa \in [-1; 1]$ ein Risikokontrollparameter,
- 10 • κ^k eine Risikokontrollfunktion $\kappa^k(\xi) = (1 - \kappa \text{sign}(\xi))\xi$,
- $\nabla J(\cdot)$ die Ableitung des Funktionsapproximators nach seinen Gewichten,
- $r(x_t, a_t, x_{t+1})$ ein Gewinn beim Zustandsübergang von dem Zustand x_t nach dem Folgezustand x_{t+1} ,

15

bezeichnet wird/werden.

Das System ist bevorzugt ein technisches System, von dem vor der Ermittlung Meßgrößen gemessen werden, die bei der Ermittlung der Folge von Aktionen verwendet werden.

Unter Verwendung der ermittelten Folge von Aktionen kann das technische System gesteuert oder auch geregelt werden.

25 Bevorzugt wird das System als ein Markov-Entscheidungsproblem modelliert.

Das Verfahren oder die Anordnung werden bevorzugt in einem Verkehrsleitsystem oder in einem Kommunikationssystem eingesetzt, wobei in dem Kommunikationssystem die Folge von Aktionen zur Durchführung einer Zugangskontrolle oder eines Routings, also einer Pfadvergabe, in einem Kommunikationsnetz eingesetzt wird.

35 Ferner kann das System ein Finanzmarkt sein, welcher durch ein Markov-Entscheidungsproblem modelliert wird und wobei der Verlauf des Finanzmarkts, beispielsweise ein Verlauf eines

Aktienindex oder auch ein Kursverlauf eines Devisenmarktes unter Verwendung der Verfahren bzw. der Anordnung analysiert und in den Markt entsprechend der Folge ermittelter Aktionen eingegriffen werden kann.

5

Ausführungsbeispiele der Erfindung sind in den Figuren dargestellt und werden im weiteren näher erläutert.

Es zeigen

10

Figur 1 ein Ablaufdiagramm, in dem einzelne Verfahrensschritte des ersten Ausführungsbeispiels dargestellt sind;

15

Figur 2 eine Skizze eines Systems, welches als Markov-Entscheidungsproblem modelliert werden kann;

Figur 3 eine Skizze eines Kommunikationsnetzes, bei dem in einer Vermittlungseinheit eine Zugangskontrolle durchgeführt wird;

20

Figur 4 eine symbolische Skizze eines Funktionsapproximators, mit dem ein Verfahren des approximativen dynamischen Programmierens realisiert wird;

25

Figur 5 eine weitere Skizze von mehreren Funktionsapproximatoren, mit dem ein approximatives dynamisches Programmieren implementiert wird;

30

Figur 6 eine Skizze eines Verkehrsleitsystems, welches gemäß einem Ausführungsbeispiel geregelt wird.

Erstes Ausführungsbeispiel: Zugangskontrolle und Routing.

Fig.3 zeigt ein Kommunikationsnetz 300, welches eine Vielzahl von Vermittlungseinheiten 301a, 301b, ..., 301i, ... 301n aufweist, die über Verbindungen 302a, 302b, 302j, ... 302m miteinander verbunden sind.

Ferner ist ein erstes Endgerät 303 mit einer ersten Vermittlungseinheit 301a verbunden. Von dem ersten Endgerät 303 wird eine Anforderungsnachricht 304 an die erste Vermittlungseinheit 301a gesendet, mit der eine Reservierung einer vorgegebenen Bandbreite innerhalb des Kommunikationsnetzes 300 zur Übertragung von Daten (Videodaten, textuelle Daten) angefordert wird.

15

In der ersten Vermittlungseinheit 301a wird gemäß einer im weiteren beschriebenen Strategie ermittelt, ob die angeforderte Bandbreite in dem Kommunikationsnetz 300 auf einer angegebenen, angeforderten Verbindung verfügbar ist (Schritt 305).

Ist dies nicht der Fall, so wird die Anforderung zurückgewiesen (Schritt 306).

25

Ist ausreichend Bandbreite verfügbar, so wird in einem weiteren Überprüfungsschritt (Schritt 307) überprüft, ob die Bandbreite reserviert werden kann.

30

Ist dies nicht der Fall, so wird die Anforderung zurückgewiesen (Schritt 308).

35

Sonst wird von der ersten Vermittlungseinheit 301a eine Route von der ersten Vermittlungseinheit 301a über weitere Vermittlungseinheiten 301i zu einem zweiten Endgerät 309, mit dem das erste Endgerät 303 kommunizieren will, ausgewählt und es wird eine Verbindung initialisiert (Schritt 310).

Im folgenden wird von einem Kommunikationsnetz 300 ausgegangen, welches einen Satz von Vermittlungseinheiten

$$N = \{1, K_1, n, K_n, N\} \quad (17)$$

5 und einen Satz von physikalischen Verbindungen

$$L = \{1, K_1, l, K_l, L\}, \quad (18)$$

umfaßt, wobei eine physikalische Verbindung l eine Kapazität
10 von $B(l)$ Bandbreiteneinheiten aufweist.

Es sind ein Satz

$$M = \{1, K_1, m, K_m, M\} \quad (19)$$

15

verschiedener Diensttypen m verfügbar, wobei ein Diensttyp m durch

• einen Bandbreitenbedarf $b(m)$,

• eine durchschnittliche Verbindungsduer $\frac{1}{v(m)}$, und

20 • einen Gewinn $c(m)$, den man dann erhält, wenn eine Verbindungsanforderung des entsprechenden Diensttyps m akzeptiert wird,

charakterisiert ist.

25 Der Gewinn $c(m)$ ist gegeben durch die Menge des Geldes, die ein Netzwerkbetreiber des Kommunikationsnetzes 300 einem Teilnehmer für eine Verbindung des Diensttyps in Rechnung stellt. Anschaulich spiegelt der Gewinn $c(m)$ unterschiedliche, von dem Netzbetreiber vorgebbare Prioritäten wider, die 30 er mit verschiedenen Diensten assoziiert.

Eine physikalischen Verbindung l kann gleichzeitig eine beliebige Kombination von Kommunikationsverbindungen bereitstellen, solange die genutzte Bandbreite der Kommunikationsverbindungen nicht die insgesamt verfügbare Bandbreite der physikalischen Verbindung übersteigt.

Wird eine neue Kommunikationsverbindung des Typs m angefordert zwischen einem ersten Knoten i und einem zweiten Knoten j (Endgeräte werden auch als Knoten bezeichnet), so kann die 5 angeforderte Kommunikationsverbindung, wie oben dargestellt, entweder akzeptiert oder zurückgewiesen werden.

Wird die Kommunikationsverbindungen akzeptiert, so wird eine Route aus einer Menge vorgegebener Routen ausgewählt. Diese 10 Auswahl wird als Routing bezeichnet. Im Rahmen der Kommunikationsverbindung vom Typ m werden $b(m)$ Bandbreiteneinheiten für jede physikalische Verbindung entlang der ausgewählten Route für die Verbindungsduer benutzt.

15 Somit kann im Rahmen der Zugangskontrolle (Call-Admission-Control) eine Route innerhalb des Kommunikationsnetzes 300 nur ausgewählt werden, wenn die ausgewählte Route ausreichend Bandbreite zur Verfügung hat.

20 Ziel der Zugangskontrolle und des Routings ist es, einen langfristigen Gewinn, der durch Akzeptanz der angeforderten Verbindungen erhalten wird, zu maximieren.

25 Das technische System Kommunikationsnetz 300 befindet sich zu einem Zeitpunkt t in einem Zustand x_t , welcher durch eine Liste von Routen über bestehende Verbindungen beschrieben wird, durch welche Listen angezeigt ist, wie viele Verbindungen welchen Diensttyps zu dem Zeitpunkt t die jeweilige Route verwenden.

30 Ereignisse w , durch die ein Zustand x_t in einen Folgezustand x_{t+1} überführt werden könnte, sind das Ankommen neuer Verbindungsanforderungsnachrichten oder auch das Beenden einer in dem Kommunikationsnetz 300 bestehenden Verbindung.

35 Eine Aktion a_t zu einem Zeitpunkt t aufgrund einer Verbindungsanforderung ist in diesem Ausführungsbeispiel die Ent-

scheidung, ob eine Verbindungsanforderung akzeptiert oder zurückgewiesen werden soll und, falls die Verbindung akzeptiert wird, die Auswahl der Route durch das Kommunikationsnetz 300.

5 Ziel ist die Ermittlung einer Folge von Aktionen, d.h. anschaulich das Lernen einer Strategie mit Aktionen zu einem Zustand x_t zu bestimmen derart, daß folgende Vorschrift maximiert wird:

$$10 \quad E \left(\sum_{k=0}^{\infty} e^{-\beta t_k} \cdot g(x_{t_k}, \omega_k, a_{t_k}) \right), \quad (20)$$

wobei mit

- $E\{\cdot\}$ ein Erwartungswert,
- 15 • t_k ein Zeitpunkt, zu dem ein k-tes Ereignis erfolgt,
- $g(x_{t_k}, \omega_k, a_{t_k})$ der Gewinn, der mit dem k-ten Ereignis verbunden ist, und
- β ein Verringerungsfaktor, der einen sofortigen Gewinn wertvoller bewertet als ein Gewinn in ferner in der Zukunft

20 liegenden Zeitpunkten,

bezeichnet wird.

Unterschiedliche Realisierungen einer Strategie führen üblicherweise zu unterschiedlichen Gesamtgewinnen G :

$$G = \sum_{k=0}^{\infty} e^{-\beta t_k} \cdot g(x_{t_k}, \omega_k, a_{t_k}). \quad (21)$$

30 Ziel ist die Maximierung des Erwartungswerts des Gesamtgewinns G gemäß folgender Vorschrift J :

$$J = E \left\{ \sum_{k=0}^{\infty} e^{-\beta t_k} \cdot g(x_{t_k}, \omega_k, a_{t_k}) \right\}, \quad (22)$$

wobei ein Risiko, daß der Gesamtgewinn G einer speziellen Realisierung einer Zugangskontrolle und einer Routing-

5 Strategie unter den Erwartungswert sinkt, einstellbar ist.

Zur Durchführung der Zugriffskontrolle sowie zum Routing wird das $TD(\lambda)$ -Lernverfahren eingesetzt.

10 Es wird folgende Zielfunktion im Rahmen dieses Ausführungsbeispiels eingesetzt:

$$J^*(x_t) = E_{\tau} \left\{ e^{-\beta \tau} \right\} E_{\omega} \left\{ \max_{a \in A} [g(x_t, \omega_t, a) + J^*(x_{t+1})] \right\}, \quad (23)$$

15 wobei mit

- A ein Aktionsraum mit einer vorgegebenen Anzahl Aktionen, die in einem Zustand x_t jeweils zur Verfügung stehen,
- τ ein erster Zeitpunkt, zudem ein erstes Ereignis ω erfolgt,
- x_{t+1} ein Folgezustand des Systems,

bezeichnet wird.

25 Ein approximierter Wert des Zielwerts $J^*(x_t)$ wird unter Verwendung eines Funktionsapproximators 400 (vgl. Fig.4) unter Verwendung von Trainingsdaten gelernt und gespeichert.

Trainingsdaten sind zuvor in dem Kommunikationsnetz 300 gemessene Daten über das Verhalten des Kommunikationsnetzes 300 bei ankommenden Verbindungsanforderungen 304 sowie bei Beendigung von Nachrichten. Diese zeitliche Folge von Zuständen wird gespeichert und mit diesen Trainingsdaten wird der Funktionsapproximator 400 gemäß dem im weiteren beschriebenen

30 Lernverfahren trainiert.

Als Eingangsgröße des Funktionsapproximators 400 dienen für jeden Eingang 401, 402, 403 des Funktionsapproximators 400 jeweils eine Anzahl von Verbindungen jeweils eines Diensttypes 5 m auf einer Route des Kommunikationsnetzes 300. Diese sind in Fig.4 durch Blöcke 404, 405, 406 symbolisch dargestellt.

Ausgangsgröße des Funktionsapproximators 400 ist ein approximierter Zielwert \tilde{J} des Zielwerts J^* .

10 Eine detailliertere Darstellung des Funktionsapproximators 500, welcher in diesem Fall mehrere Teil-Funktionsapproximatoren 510, 520 des Funktionsapproximators 500 aufweist, zeigt Fig.5. Eine Ausgangsgröße ist der 15 approximierte Zielwert \tilde{J} , der gemäß folgender Vorschrift gebildet wird:

$$\tilde{J}(x_t, \Theta) = \sum_{l=1}^L \tilde{J}^{(l)}(x_t^{(l)}, \Theta_t^{(l)}). \quad (24)$$

20 Die Eingangsgrößen der Teilfunktionsapproximatoren 510, 520, die an Eingängen 511, 512, 513 des ersten Teil-Funktionsapproximators 510 bzw. an Eingängen 521, 522 und 523 des zweiten Teilfunktionsapproximators 520 anliegen, sind wiederum jeweils eine Anzahl von Diensttypen eines Typs m jeweils in einer physikalischen Verbindung r, symbolisiert 25 durch Blöcke 514, 515, 516 für den ersten Teil-Funktionsapproximator und 524, 525 und 526 für den zweiten Teil-Funktionsapproximator 520.

30 Teilausgangsgrößen 530, 531, 532, 533 werden einer Addiereinheit 540 zugeführt und als Ausgangsgröße der Addiereinheit wird die approximierte Zielgröße \tilde{J} gebildet.

35 Angenommen, das Kommunikationsnetz 300 befindet sich in dem Zustand x_{t_k} und eine Anforderungsnachricht, mit der ein Diensttyp m der Klasse m für eine Verbindung zwischen zwei

Knoten i, j angefordert wird, gelangt zu der ersten Verbindungseinheit 301a.

5 Mit $R(i, j)$ wird eine Liste erlaubter Routen zwischen den Knoten i und j bezeichnet und mit

$$\tilde{R}(i, j, x_{t_k}) \subset R(i, j) \quad (25)$$

10 wird eine Liste aller möglichen Routen als Teilmenge der Routen $R(i, j)$ bezeichnet, die hinsichtlich der verfügbaren und angeforderten Bandbreite eine mögliche Verbindung realisieren könnten.

15 Für jede mögliche Route $r, r \in \tilde{R}(i, j, x_{t_k})$ wird ein Folgezustand $x_{t_k+1}(x_{t_k}, \omega_k, r)$ ermittelt, der daraus resultiert, daß die Verbindungsanforderung 304 akzeptiert wird und die Verbindung auf der Route r dem anfordernden ersten Endgerät 303 zur Verfügung gestellt wird.

20 Dies ist in Fig.1 als zweiter Schritt (Schritt 102) dargestellt, wobei in einem ersten Schritt (Schritt 101) jeweils der Zustand des Systems sowie das jeweilige Ereignis festgestellt werden.

25 Es wird in einem dritten Schritt (Schritt 103) eine auszuwählende Route r^* gemäß folgender Vorschrift ermittelt:

$$r^* = \arg \max_{r \in \tilde{R}(i, j, x_{t_k})} \tilde{J}(x_{t_k+1}(x_{t_k}, \omega_k, r), \Theta_t). \quad (26)$$

30 In einem weiteren Schritt (Schritt 104) wird überprüft, ob folgende Vorschrift erfüllt ist:

$$c(m) + \tilde{J}(x_{t_k+1}(x_{t_k}, \omega_k, r^*), \Theta_t) < \tilde{J}(x_{t_k}, \Theta_t). \quad (27)$$

Ist dies der Fall, so wird die Verbindungsanforderung 304 zurückgewiesen (Schritt 105), sonst wird die Verbindung akzeptiert und entlang der ausgewählten Route r^* zu dem Knoten j „durchgeschaltet“ (Schritt 106).

5

In einem Parametervektor Θ sind jeweils für einen Zeitpunkt t Gewichte des Funktionsapproximators 400, 500 gespeichert, die im Rahmen des $TD(\lambda)$ -Lernverfahrens an die Trainingsdaten adaptiert werden, so daß eine optimierte Zugangskontrolle und 10 ein optimiertes Routing erreicht wird.

Während der Trainingsphase werden die Gewichtsparameter an die dem Funktionsapproximator angelegten Trainingsdaten angepaßt.

15

Es wird ein Risikoparameter κ definiert, mit dem ein gewünschtes Risiko, welches durch eine Folge von Aktionen und Zuständen hinsichtlich eines vorgegebenen Zustands des Systems aufweist, einstellbar ist, gemäß folgenden Vorschriften:

20

$-1 \leq \kappa < 0$: risikoreiches Lernen,

$\kappa = 0$: hinsichtlich des Risikos ein neutrales Lernen,

$0 < \kappa < 1$: ein risiko-vermeidendes Lernen,

$\kappa = 1$: „Worst-Case“-Lernen.

30 Fernen wird im Rahmen des Lernverfahrens ein vorgebbarer Parameter $0 \leq \lambda \leq 1$ und eine Schrittweitenfolge γ_k vorgegeben.

Die Gewichtswerte des Gewichtsvektors Θ werden aufgrund jedes Ereignisses ω_{t_k} gemäß folgender Adoptionsvorschrift an

35 die Trainingsdaten angepaßt:

$$\Theta_k = \Theta_{k-1} + \gamma_k N^k(d_k) z_t, \quad (28)$$

wobei

$$d_k = e^{-\beta(t_k - t_{k-1})} (g(x_{t_k}, \omega_k, a_{t_k}) + \tilde{J}(x_{t_k}, \Theta_{k-1})) - \tilde{J}(x_{t_{k-1}}, \Theta_{k-1}) \quad (29)$$

5

$$z_t = \lambda e^{-\beta(t_{k-1} - t_{k-2})} z_{t-1} + \nabla_{\Theta} \tilde{J}(x_{t_{k-1}}, \Theta_{k-1}), \quad (30)$$

und

10

$$\kappa^k(\xi) = (1 - \kappa \text{sign}(\xi))\xi. \quad (31)$$

Es wird angenommen: $z_{-1} = 0$.

15 Die Funktion

$$g(x_{t_k}, \omega_k, a_{t_k}) \quad (32)$$

bezeichnet den sofortigen Gewinn gemäß folgender Vorschrift:

20

$$g(x_{t_k}, \omega_k, a_{t_k}) = \begin{cases} c(m) & \text{wenn } \omega_{t_k} \text{ ist eine Dienstanforderung eines} \\ & \text{Diensttypes } m \text{ und die Verbindung wird} \\ & \text{akzeptiert} \\ 0 & \text{sonst} \end{cases} \quad (33)$$

Es wird also, wie oben beschrieben, eine Folge von Aktionen
25 ermittelt, hinsichtlich einer Verbindungsanforderung, so daß
eine Verbindungsanforderung aufgrund einer Aktion entweder
zurückgewiesen oder akzeptiert wird. Die Ermittlung erfolgt
unter Berücksichtigung einer Optimierungsfunktion, in der das
Risiko mittels eines Risikokontrollparameters $\kappa \in [-1; 1]$
30 variabel einstellbar ist.

Zweites Ausführungsbeispiel: Verkehrsleitsystem

Fig. 6 zeigt eine Straße 600, die von Autos 601, 602, 603, 604, 605 und 606 befahren ist.

5

In die Straße 600 integrierte Leiterschleifen 610, 611 nehmen elektrische Signale in bekannter Weise auf und führen die elektrischen Signale 615, 616, einem Rechner 620 über eine Eingangs-/Ausgangsschnittstelle 621 zu. In einem mit der Eingangs-/Ausgangsschnittstelle 621 verbundenen Analog-/Digital-Wandler 622 werden die elektrischen Signale in eine Zeitreihe digitalisiert und in einem Speicher 623, der über einen Bus 624 mit dem Analog-/Digital-Wandler 622 und einem Prozessor 625 verbunden ist, gespeichert. Über die Eingangs-/Ausgangsschnittstelle 621 werden einem Verkehrsleitsystem 650 Steuerungssignale 651 zugeführt, aus denen in dem Verkehrsleitsystem 650 eine vorgegebene Geschwindigkeitsvorgabe 652 einstellbar ist oder auch weitere Angaben von Verkehrs- vorschriften, die über das Verkehrsleitsystem 650 Fahrern der Fahrzeuge 601, 602, 603, 604, 605 und 606 dargestellt werden.

Zur Verkehrsmodellierung werden in diesem Fall folgende lokale Zustandsgrößen verwendet:

- Verkehrsflußgeschwindigkeit v ,
- 25 • Fahrzeugdichte ρ (ρ = Anzahl von Fahrzeugen pro Kilometer $\frac{Fz}{km}$),
- Verkehrsfluß q (q = Anzahl der Fahrzeuge pro Stunde $\frac{Fz}{h}$,
 $(q = v * \rho)$), und
- 30 • jeweils zu einem Zeitpunkt von dem Verkehrsleitsystem 650 angezeigte Geschwindigkeitsbegrenzungen 652.

Die lokalen Zustandsgrößen werden wie oben beschrieben unter Verwendung der Leiterschleifen 610, 611 gemessen.

Somit stellen diese Größen ($v(t)$, $p(t)$, $q(t)$) einen Zustand des technischen Systems "Verkehr" zu einem bestimmten Zeitpunkt t dar.

5 In diesem Ausführungsbeispiel ist somit das System ein Verkehrssystem, welches unter Verwendung des Verkehrsleitsystems 650 geregelt wird.

10 Als Verfahren des approximativen dynamischen Programmierens wird in diesem zweiten Ausführungsbeispiel ein erweitertes Q-Lernverfahren beschrieben.

Der Zustand x_t wird beschrieben durch einen Zustandsvektor

15 $x(t) = (v(t), p(t), q(t))$. (34)

Die Aktion a_t bezeichnet die Geschwindigkeitsbegrenzung 652, die zum Zeitpunkt t von dem Verkehrsleitsystem 650 angezeigt wird.

20 Der Gewinn $r(x_t, a_t, x_{t+1})$ beschreibt die Güte des Verkehrsflusses, der zwischen den Zeitpunkten t und $t+1$ von den Leiterschleifen 610 und 611 gemessen wurde. Im Rahmen dieses zweiten Ausführungsbeispiels bezeichnet $r(x_t, a_t, x_{t+1})$

25

- die mittlere Geschwindigkeit der Fahrzeuge im Zeitintervall $[t, t + 1]$,

oder

30

- die Anzahl der Fahrzeuge, die im Zeitintervall $[t, t + 1]$ die Leiterschleifen 610 und 611 passiert haben,

oder

35

- die Varianz der Fahrzeuggeschwindigkeiten im Zeitintervall $[t, t + 1]$,

oder

- eine gewichtete Summe aus den obigen Größen.

5

Für jede mögliche Aktion a_t , d.h. für jede von dem Verkehrsleitsystem 650 anzeigbare Geschwindigkeitsbegrenzung, wird ein Wert der Optimierungsfunktion OFQ ermittelt, wobei jeweils ein geschätzter Wert der Optimierungsfunktion OFQ als

10 neuronales Netz realisiert wird.

Aus dieser Weise ergibt sich eine Menge von Bewertungsgrößen für die unterschiedlichen Aktionen a_t in dem Systemzustand x_t .

15

In einer Regelungsphase wird aus den möglichen Aktionen a_t , d.h. aus der Menge der von dem Verkehrsleitsystem 650 anzeigbaren Geschwindigkeitsbegrenzungen, diejenige Aktion a_t ausgewählt, für die in dem aktuellen Systemzustand x_t die maximale Bewertungsgröße OFQ ermittelt worden ist.

20

Die aus dem Q-Lernverfahren bekannte Adoptionsvorschrift zur Berechnung der Optimierungsfunktion OFQ wird gemäß diesem Ausführungsbeispiel um eine Risikokontrollfunktion $\kappa^k(\cdot)$, die 25 das Risiko berücksichtigt, erweitert.

Wiederum wird der Risikokontrollparameter κ gemäß der Strategie aus dem ersten Ausführungsbeispiel im Intervall von $[-1 \leq \kappa \leq 1]$ vorgegeben und repräsentiert das Risiko, das ein 30 Benutzer im Rahmen der Anwendung hinsichtlich der zu bestimmenden Kontrollstrategie eingehen will.

Gemäß diesem Ausführungsbeispiel wird folgende Bewertungsfunktion OFQ verwendet:

35

$$OFQ = Q(x; w^a), \quad (35)$$

wobei mit

- $x = (v; \rho; q)$ ein Zustand des Verkehrssystems,
- 5 • a eine Geschwindigkeitsbegrenzung aus dem Aktionsraum A aller vom Verkehrsleitsystem 650 anzeigbaren Geschwindigkeitsbegrenzungen,
- w^a die zur Geschwindigkeitsbegrenzung a gehörigen Gewichte des neuronalen Netzes,

10 bezeichnet wird/werden.

Im Rahmen des Q-Lernens wird zur Ermittlung der optimalen Gewichte w^a den neuronalen Netzes folgender Adaptionsschritt

15 ausgeführt:

$$w_{t+1}^{at} = w_t^{at} + \eta_t \cdot \mathbf{N}^k(d_t) \cdot \nabla Q(x_t; w_t^{at}) \quad (36)$$

mit der Abkürzung

$$d_t = r(x_t, a_t, x_{t+1}) + \gamma \max_{a \in A} Q(x_{t+1}, w_t^a) - Q(x_t, w_t^{at}) \quad (37)$$

wobei mit

- 25 • x_t, x_{t+1} jeweils ein Zustand des Verkehrssystems gemäß Vorschrift (34),
- a_t eine Aktion, d.h. eine von dem Verkehrsleitsystem 650 anzeigbare Geschwindigkeitsbegrenzung,
- γ ein vorgebbarer Verringerungsfaktor,
- 30 • w_t^{at} der zur Aktion a_t gehörige Gewichtsvektor vor dem Adaptionsschritt,
- w_{t+1}^{at} der zur Aktion a_t gehörige Gewichtsvektor nach dem Adaptionsschritt,
- η_t ($t = 1, \dots$) eine vorgebbare Schrittweitenfolge,

- $\kappa \in [-1; 1]$ ein Risikokontrollparameter,
- κ^k eine Risikokontrollfunktion $\kappa^k(\xi) = (1 - \kappa \text{sign}(\xi))\xi$,
- $\nabla Q(\cdot)$ die Ableitung des neuronalen Netzes nach seinen Gewichten,
- 5 • $r(x_t, a_t, x_{t+1})$ ein Gewinn beim Zustandsübergang von dem Zustand x_t nach dem Folgezustand x_{t+1} ,

bezeichnet wird/werden.

10 Im Rahmen des Lernens kann aus den möglichen Aktionen a_t eine Aktion a_t zufällig gewählt werden. Hierbei ist es nicht erforderlich, die Aktion a_t , die zu der größten Bewertungsgröße geführt hat, zu wählen.

15 Die Adaption der Gewichte hat derart zu erfolgen, daß nicht nur eine im Erwartungswert der Optimierungsfunktion optimierte Regelung des Verkehrs erreicht wird, sondern auch eine Varianz der Regelungsergebnisse berücksichtigt wird.

20 Dies ist insbesondere vorteilhaft, da der Zustandsvektor $x(t)$ das tatsächliche System Verkehr in einigen Aspekten nur unzureichend modelliert und es deshalb zu nicht erwarteten Störungen kommen kann. So hängt die Dynamik des Verkehrs und damit seiner Modellierung von weiteren Faktoren wie beispielsweise Wetter, Anteil an Lastkraftwagen auf der Straße, ein Anteil von Wohnmobilen, etc. ab, die nicht immer in den Meßgrößen des Zustandsvektors $x(t)$ integriert sind. Zudem ist nicht immer sichergestellt, daß die Verkehrsteilnehmer sofort den neuen Geschwindigkeitsangaben gemäß dem Verkehrsleitsystem Folge leisten.

25

30

Eine Regelungsphase an dem realen System gemäß dem Verkehrsleitsystem vollzieht sich gemäß folgenden Schritten:

35 1. Das Messen des Zustandes x_t zum Zeitpunkt t erfolgt an verschiedenen Stellen des Verkehrssystems Verkehr und ergibt einen Zustandsvektor $x(t) = (v(t), \rho(t), q(t))$.

2. Für alle möglichen Aktionen a_t wird ein Wert der Optimierungsfunktion ermittelt und es wird diejenige Aktion a_t mit der höchsten Bewertung in der Optimierungsfunktion ausgewählt.

In diesem Dokument sind folgende Veröffentlichungen zitiert:

- 5 [1] R. Neuneier, Enhancing Q-Learning for Optimal Asset Al-
location, Proceedings of the Neural Information Proces-
sing Systems, NIPS 1997
- [2] R.S. Sutton, Learning to predict by the method of tempo-
ral differences, Machine Learning, 3:9-44, 1988
- 10 [3] M. Heger, Risk and Reinforcement Learning: Concepts and
Dynamic Programming, ZKW Bericht Nr. 8/94, Zentrum für
Kognitionswissenschaften, Universität Bremen, ISSN 0947-
0204, Dezember 1994
- 15 [4] D.P. Bertsekas, Dynamic Programming and Optimal Control,
Athena Scientific, Belmont, MA, 1995

Patentansprüche

1. Verfahren zur rechnergestützten Ermittlung einer Folge von Aktionen für ein System, welches Zustände aufweist, wobei ein Zustandsübergang zwischen zwei Zuständen aufgrund einer Aktion erfolgt,
5 bei dem die Ermittlung der Folge von Aktionen derart erfolgt, daß eine aus der Folge von Aktionen resultierende Folge von Zuständen hinsichtlich einer vorgegebenen Optimierungsfunktion optimiert ist, wobei die Optimierungsfunktion einen variablen Parameter enthält, mit dem ein Risiko, welches die resultierende Folge von Zuständen hinsichtlich eines vorgegebenen Zustandes des Systems aufweist, einstellbar ist.
- 10 15 2. Verfahren nach Anspruch 1,
bei dem zur Ermittlung ein Verfahren des approximativen Dynamischen Programmierens eingesetzt wird.
- 20 3. Verfahren nach Anspruch 2,
bei dem das Verfahren des approximativen Dynamischen Programmierens ein auf dem Q-Lernen basierendes Verfahren ist.
- 25 4. Verfahren nach Anspruch 3,
bei dem die Optimierungsfunktion OFQ im Rahmen des Q-Lernens gemäß folgender Vorschrift gebildet wird:

$$OFQ = Q(x; w^a),$$

wobei mit

- 30 35 • x ein Zustand in einem Zustandsraum X ,
- a eine Aktion aus einem Aktionsraum A ,
- w^a die zur Aktion a gehörigen Gewichte eines Funktionsapproximators

bezeichnet wird/werden, und bei dem die Gewichte des Funktionsapproximators gemäß folgender Vorschrift adaptiert werden:

$$w_{t+1}^{a_t} = w_t^{a_t} + \eta_t \cdot \kappa^\kappa(d_t) \cdot \nabla Q(x_t; w_t^{a_t})$$

mit der Abkürzung

5

$$d_t = r(x_t, a_t, x_{t+1}) + \gamma \max_{a \in A} Q(x_{t+1}, w_t^a) - Q(x_t, w_t^{a_t})$$

wobei mit

10 • x_t, x_{t+1} jeweils ein Zustand in dem Zustandsraum X ,
 • a_t eine Aktion aus einem Aktionsraum A ,
 • γ ein vorgebbarer Verringerungsfaktor,
 • $w_t^{a_t}$ der zur Aktion a_t gehörige Gewichtsvektor vor dem Adaptionsschritt,
 15 • $w_{t+1}^{a_t}$ der zur Aktion a_t gehörige Gewichtsvektor nach dem Adaptionsschritt,
 • η_t ($t = 1, \dots$) eine vorgebbare Schrittweitenfolge,
 • $\kappa \in [-1; 1]$ ein Risikokontrollparameter,
 • κ^κ eine Risikokontrollfunktion $\kappa^\kappa(\xi) = (1 - \kappa \text{sign}(\xi))\xi$,
 20 • $\nabla Q(\cdot)$ die Ableitung des Funktionsapproximators nach seinen Gewichten,
 • $r(x_t, a_t, x_{t+1})$ ein Gewinn beim Zustandsübergang von dem Zustand x_t nach dem Folgezustand x_{t+1} ,
 25 bezeichnet wird/werden.

5. Verfahren nach Anspruch 2,

bei dem das Verfahren des approximativen Dynamischen Programmierens ein auf dem $TD(\lambda)$ -Lernen basierendes Verfahren ist.

30

6. Verfahren nach Anspruch 5,

bei dem die Optimierungsfunktion OFTD im Rahmen des $TD(\lambda)$ -Lernens gemäß folgender Vorschrift gebildet wird:

$$\text{OFTD} = J(x; w)$$

wobei mit

5 • x ein Zustand in einem Zustandsraum X ,
 • a eine Aktion aus einem Aktionsraum A ,
 • w die Gewichte eines Funktionsapproximators

10 bezeichnet wird/werden, und bei dem die Gewichte des Funkti-
 onsapproximators gemäß folgender Vorschrift adaptiert werden:

$$w_{t+1} = w_t + \eta_t \cdot \nabla J(d_t) \cdot z_t$$

mit den Abkürzungen

15 $d_t = r(w_t, a_t, x_{t+1}) + \gamma J(x_{t+1}; w_t) - J(x_t; w_t),$

$$z_t = \lambda \cdot \gamma \cdot z_{t-1} + \nabla J(x_t; w_t),$$

20 $z_{-1} = 0,$

wobei mit

• x_t, x_{t+1} jeweils ein Zustand in dem Zustandsraum X ,
 25 • a_t eine Aktion aus einem Aktionsraum A ,
 • γ ein vorgebbarer Verringerungsfaktor,
 • w_t der Gewichtsvektor vor dem Adaptionsschritt,
 • w_{t+1} der Gewichtsvektor nach dem Adaptionsschritt,
 • η_t ($t = 1, \dots$) eine vorgebbare Schrittweitenfolge,
 30 • $\kappa \in [-1; 1]$ ein Risikokontrollparameter,
 • $\nabla J(\cdot; \cdot)$ die Ableitung des Funktionsapproximators nach seinen
 Gewichten,
 • $r(x_t, a_t, x_{t+1})$ ein Gewinn beim Zustandsübergang von dem Zu-
 35 stand x_t nach dem Folgezustand x_{t+1} ,

bezeichnet wird/werden.

7. Verfahren nach einem der Ansprüche 1 bis 6,
bei dem das System ein technisches System ist, von dem vor
5 der Ermittlung Meßgrößen gemessen werden, die bei der Ermitt-
lung der Folge von Aktionen verwendet werden.

8. Verfahren nach Anspruch 7,
bei dem gemäß der Folge von Aktionen das technische System
10 gesteuert wird.

9. Verfahren nach Anspruch 7,
bei dem gemäß der Folge von Aktionen das technische System
geregelt wird.

15 10. Verfahren nach einem der Ansprüche 1 bis 9,
bei dem das System als ein Markov-Entscheidungsproblem modelliert wird.

20 11. Verfahren nach einem der Ansprüche 1 bis 10,
eingesetzt in einem Verkehrsleitsystem.

12. Verfahren nach einem der Ansprüche 1 bis 10,
eingesetzt in einem Kommunikationssystem.

25 13. Verfahren nach einem der Ansprüche 1 bis 10,
eingesetzt zur Durchführung einer Zugangskontrolle in einem
Kommunikationsnetz.

30 14. Verfahren nach einem der Ansprüche 1 bis 10,
eingesetzt zur Durchführung eines Routings in einem Kommunikationsnetz.

15. Anordnung zur Ermittlung einer Folge von Aktionen für ein
35 System, welches Zustände aufweist, wobei ein Zustandsübergang
zwischen zwei Zuständen aufgrund einer Aktion erfolgt,

mit einem Prozessor, der derart eingerichtet ist, daß die Ermittlung der Folge von Aktionen derart erfolgen kann, daß eine aus der Folge von Aktionen resultierende Folge von Zuständen hinsichtlich einer vorgegebenen Optimierungsfunktion optimiert ist, wobei die Optimierungsfunktion einen variablen Parameter enthält, mit dem ein Risiko, welches die resultierende Folge von Zuständen hinsichtlich eines vorgegebenen Zustandes des Systems aufweist, einstellbar ist.

10 16. Anordnung nach Anspruch 15,
eingesetzt zur Steuerung eines technischen Systems.

17. Anordnung nach Anspruch 15,
eingesetzt zur Regelung eines technischen Systems.

15 18. Anordnung nach Anspruch 15,
eingesetzt in einem Verkehrsleitsystem.

19. Anordnung nach Anspruch 15,
eingesetzt in einem Kommunikationssystem.

20. Anordnung nach Anspruch 15,
eingesetzt zur Durchführung einer Zugangskontrolle in einem Kommunikationsnetz.

25 21. Anordnung nach Anspruch 15,
eingesetzt zur Durchführung eines Routings in einem Kommunikationsnetz.

Zusammenfassung

5 **Verfahren und Anordnung zur Ermittlung einer Folge von Aktionen für ein System, welches Zustände aufweist, wobei ein Zustandsübergang zwischen zwei Zuständen aufgrund einer Aktion erfolgt**

10 Die Ermittlung der Folge von Aktionen erfolgt derart, daß eine aus der Folge von Aktionen resultierende Folge von Zuständen hinsichtlich einer vorgegebenen Optimierungsfunktion optimiert ist. Die Optimierungsfunktion enthält einen variablen Parameter, mit dem ein Risiko einstellbar ist, welches Risiko die resultierende Folge von Zuständen hinsichtlich eines vorgegebenen Zustandes des Systems aufweist.

FIG 1

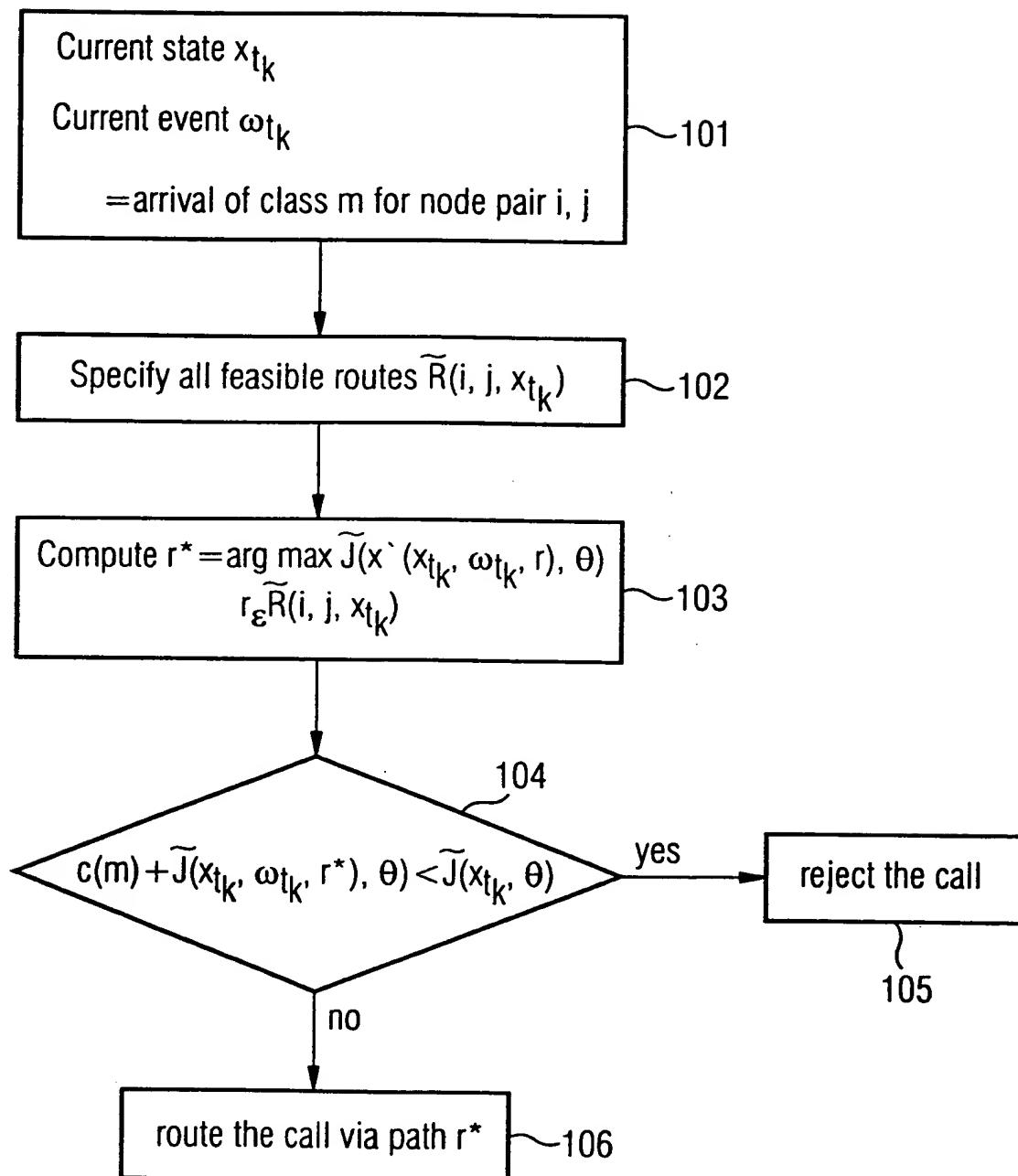


FIG 2

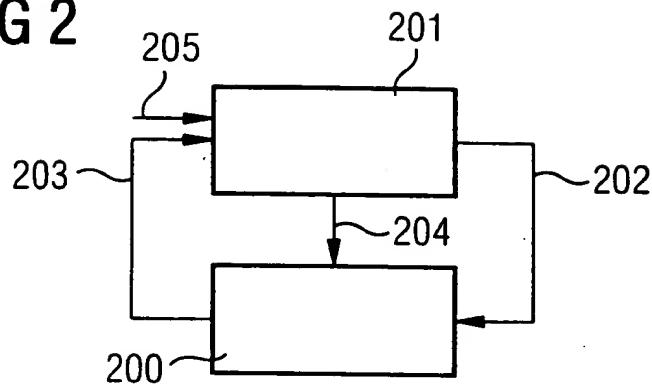


FIG3

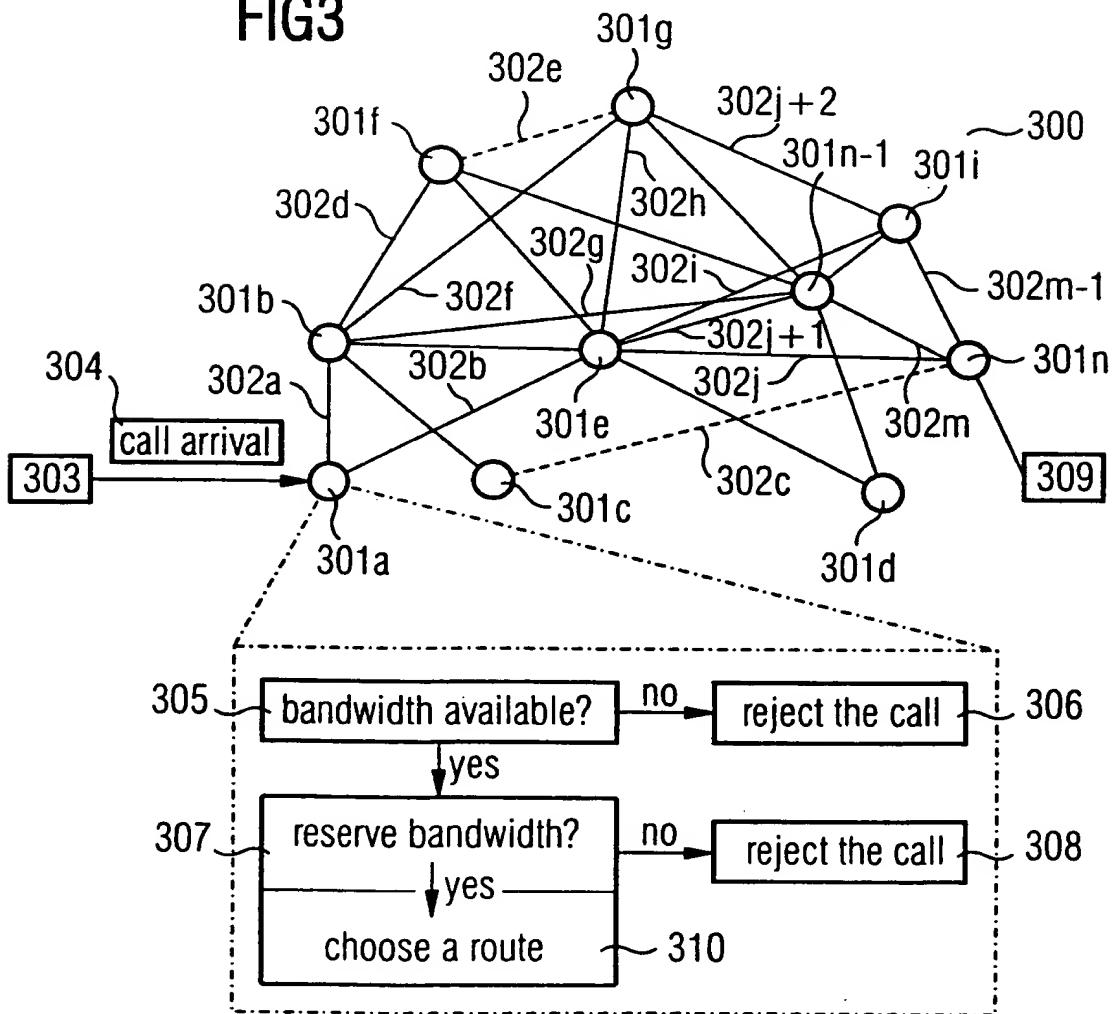


FIG 4

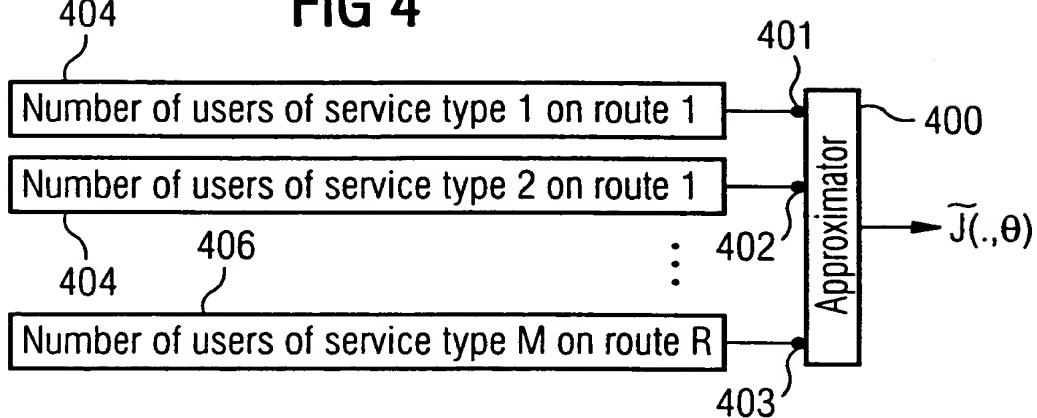


FIG 5

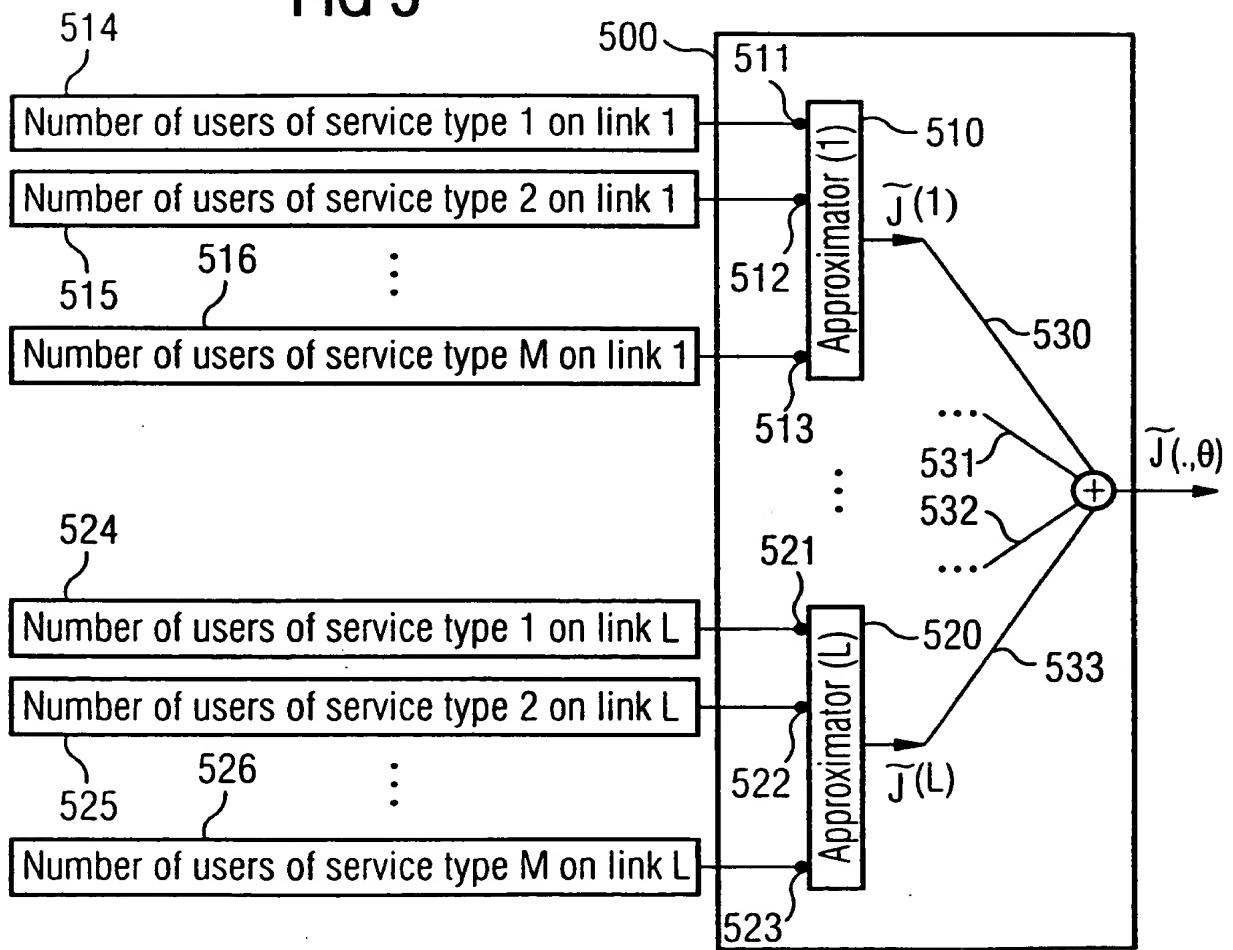


FIG 6

